(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) N° de publication : **2 799 024**
(à n'utiliser que pour les commandes de reproduction)

(21) N° d'enregistrement national : **99 12141**

(51) Int Cl[7] : G 06 F 17/30

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt** : 29.09.99.

(30) **Priorité** :

(43) **Date de mise à la disposition du public de la demande** : 30.03.01 Bulletin 01/13.

(56) **Liste des documents cités dans le rapport de recherche préliminaire** : *Se reporter à la fin du présent fascicule*

(60) **Références à d'autres documents nationaux apparentés** :

(71) **Demandeur(s)** : *OBJECTMINE Société à responsabilité limitée* — FR.

(72) **Inventeur(s)** : BREGEAULT LUC.

(73) **Titulaire(s)** :

(74) **Mandataire(s)** : REGIMBEAU.

(54) SYSTEME ET PROCEDE DE TRAITEMENT DE DONNEES DESTINEES A ENRICHIR LES SYSTEMES D'AIDE A LA DECISION.

(57) Un système de traitement de données comprend:
- un serveur contenant dans une mémoire un jeu de données individuelles organisées sous forme d'au moins une variable à N dimensions, chaque donnée individuelle contenant au moins une variable individuelle en association avec des valeurs prédéfinies de paramètres selon les N dimensions, et
- au moins un poste client communiquant avec le serveur via un espace de travail associé au serveur.

Le serveur place dans ledit espace de travail des données constituées par des combinaisons prédéfinies desdites variables individuelles à partir d'un filtrage sur lesdits paramètres, réalisé selon au moins une dimension choisie par un utilisateur parmi les N dimensions.

Selon l'invention, il est prévu à l'extérieur du serveur des moyens pour constituer temporairement des paramètres additionnels à partir d'un traitement d'analyse sur des combinaisons choisies de variables placées dans l'espace de travail, et des moyens pour élaborer dans ledit espace de travail de nouvelles combinaisons desdites variables à partir d'un nouveau filtrage mettant en jeu lesdits paramètres additionnels.

FR 2 799 024 - A1

La présente invention concerne d'une façon générale les systèmes informatiques de bases de données, et plus précisément les systèmes d'aide à la décision notamment de type à traitement analytique en ligne (OLAP pour « On Line Analytical Processing ».

De tels systèmes permettent de stocker, de gérer et de visualiser des données sous la forme de tableaux croisés dynamiques multidimensionnels.

Offrant des outils puissants de construction et de suivi de tableaux de bord, ils constituent de nous jours une aide importante à la décision dans le pilotage de l'activité économique et industrielle d'une entreprise.

Ces outils décisionnels sont destinés à coopérer avec des « entrepôts de données » (« data warehouse ») qui rassemblent l'ensemble des données gérées par l'entreprise. Ces entrepôts viennent nourrir le décisionnel, en lui fournissant des types très divers de données agrégées de l'activité. Dans une configuration classique, ces entrepôts de données et les outils associés sont contenus dans un serveur auquel accèdent une série de postes clients.

La mise en place d'un environnement décisionnel nécessite la définition des données de pilotage (dites variable métier) qui sont définies et configurées dans le serveur au moment de l'installation du système dans l'entreprise. A l'issue de la phase de configuration, la structure du système décisionnel est figée pour les utilisateurs. Elle peut néanmoins être adaptée et étendue à la demande par l'administrateur du système.

Sur le poste client, l'utilisateur va pouvoir naviguer et visualiser graphiquement ses données métier au travers de tableaux de bord graphiques pré-configurés.

Un tableau de bord permet la visualisation d'une ou plusieurs variables métier suivant des axes de représentation prédéfinis (appelés chacun « dimension »).

On comprend que, du fait que les traitements effectués par le système décisionnel sont figés, et ne peuvent être modifiés que de façon centralisée par l'administrateur du serveur, un tel système souffre d'un manque de souplesse. En particulier, lorsqu'un utilisateur souhaite créer un nouvel axe (nouvelle dimension) de visualisation non prédéfini dans le système, il doit alors faire appel à

l'administrateur qui, si la demande des différents utilisateurs est suffisamment cohérente, va alors faire effectuer les modifications nécessaires dans le serveur.

La présente invention vise, à partir d'un entrepôt de données dont la structure est définie, et d'un système décisionnel dont les fonctions sont également définies, à permettre d'offrir à l'utilisateur des visions additionnelles des données, sans qu'il soit nécessaire de modifier l'environnement applicatif, le serveur de données ou les données elles-mêmes, et donc d'en modifier la cohérence et l'intégrité.

Plus particulièrement, la présente invention vise à étendre par de nouveaux axes d'analyse une ou plusieurs variables métier étudiées par un utilisateur, et ce de façon dynamique et en temps réel.

L'invention propose à cet effet un système de traitement de données, comprenant :

- un serveur contenant dans une mémoire un jeu de données individuelles organisées sous forme d'au moins une variable à N dimensions, chaque donnée individuelle contenant au moins une variable individuelle en association avec des valeurs prédéfinies de paramètres selon les N dimensions, et

- au moins un poste client communiquant avec le serveur via un espace de travail associé au serveur,

le serveur étant apte à placer dans ledit espace de travail des données constituées par des combinaisons prédéfinies desdites variables individuelles à partir d'un filtrage sur lesdits paramètres, réalisé selon au moins une dimension choisie par un utilisateur parmi les N dimensions,

système caractérisé en ce qu'il comprend également, à l'extérieur du serveur, des moyens pour constituer temporairement des paramètres additionnels à partir d'un traitement d'analyse sur des combinaisons choisies de variables placées dans l'espace de travail, et des moyens pour élaborer dans ledit espace de travail de nouvelles combinaisons desdites variables à partir d'un nouveau filtrage mettant en jeu lesdits paramètres additionnels.

Des aspects préférés, mais non limitatifs, du système selon l'invention sont les suivants :

- les moyens pour constituer des paramètres additionnels comprennent des moyens pour établir des catégories selon un processus de catégorisation sur lesdites combinaisons de variables, lesdits paramètres additionnels étant constitués par lesdites catégories.

- les moyens pour constituer des paramètres additionnels comprennent des moyens pour établir des catégories selon un processus de marquage sur lesdites combinaisons de variables, lesdits paramètres additionnels étant constitués par des combinaisons de paramètres existants à partir desquelles un filtrage sur les données est effectué.

- les moyens pour élaborer de nouvelles combinaisons desdites variables comprennent des moyens pour pré-calculer et stocker lesdites nouvelles combinaisons.

- les moyens pour élaborer de nouvelles combinaisons desdites variables comprennent des moyens pour calculer dynamiquement lesdites nouvelles combinaisons à partir d'un filtrage appliqué au serveur et mettant en jeu lesdits paramètres additionnels.

Avantageusement, le système est utilisé dans un système décisionnel par traitement analytique en ligne.

Selon un deuxième aspect, la présente invention propose un procédé de traitement de données et de visualisation desdites données sur un poste client à partir d'un serveur contenant dans une mémoire un jeu de données individuelles organisées sous forme d'au moins une variable à N dimensions, chaque donnée individuelle contenant au moins une variable individuelle en association avec des valeurs prédéfinies de paramètres selon les N dimensions, caractérisé en ce qu'il comprend les étapes consistant à :

- fournir au poste client, via un espace de travail associé au serveur, des données constituées par des combinaisons prédéfinies desdites variables individuelles à partir d'un filtrage sur lesdits paramètres, réalisé selon au moins une dimension choisie par un utilisateur parmi les N dimensions,

- dans l'espace de travail, constituer temporairement des paramètres additionnels à partir d'un traitement d'analyse effectué sur des combinaisons choisies desdites variables, et

- dans l'espace de travail, élaborer de nouvelles combinaisons desdites variables à partir d'un nouveau filtrage mettant en jeu lesdits paramètres additionnels.

Des aspects préférés, mais non limitatifs, du procédé de traitement selon l'invention sont les suivants :

- ledit traitement d'analyse est un traitement de catégorisation sur lesdites combinaisons de variables, lesdits paramètres additionnels étant constitués par lesdites catégories.

- ledit traitement d'analyse comprend un processus de marquage, lesdits paramètres additionnels étant constitués par des combinaisons de paramètres existants à partir desquelles un filtrage sur les données du serveur est effectué.

- l'étape d'élaboration de nouvelles combinaisons desdites variables comprend un pré-calcul et un stockage desdites combinaisons dans l'espace de travail.

- l'étape d'élaboration de nouvelles combinaisons desdites variables comprend un calcul dynamique desdites combinaisons à partir d'un filtrage sur les données du serveur mettant en jeu lesdits paramètres additionnels.

D'autres aspects, buts et avantages de la présente invention apparaîtront mieux à la lecture de la description détaillée suivante de formes de réalisation préférées de celle-ci, donnée à titre d'exemple non limitatif et faite en référence aux dessins annexés, sur lesquels :

la figure 1 montre les premières lignes d'une base de données ou entrepôt de données utilisée à titre d'exemple non limitatif pour illustrer l'invention,

la figure 2 illustre un graphique pré-paramétré pouvant être engendré avec un système décisionnel classique à partir de données du type de celles de la figure 1,

la figure 3 illustre graphiquement des subdivisions selon une dimension existante et selon une nouvelle dimension de la population de la base de données,

les figures 4a et 4b illustrent respectivement un graphique pré-paramétré engendré avec le système décisionnel classique et un graphique pouvant être obtenu avec la présente invention, et

les figures 5a et 5b illustrent respectivement un autre graphique pré-paramétré engendré avec le système décisionnel classique et un autre graphique pouvant être obtenu avec la présente invention.

On va maintenant décrire en référence aux dessins différentes formes de réalisations possibles de l'invention.

1) Structure des données

Les données d'un entrepôt de données destiné à être exploité par un système décisionnel sont stockées dans des variables auxquelles sont associées des dimensions. Ces variables peuvent être représentées sous la forme :

$$V = f(D1, D2, \ldots, DN)$$

où

V est la variable qui stocke la valeur de la donnée (par exemple une variable VENTES)

Di est une dimension de représentation de la variable (par exemple une dimension TYPE DE PRODUIT)

f est une fonction de calcul des agrégats (typiquement des fonctions telles que SOMME, VALEUR MOYENNE, MAXIMUM, MINIMUM, etc.).

Une dimension est composée d'éléments pouvant être en général organisés de façon hiérarchique. Les feuilles (c'est-à-dire les éléments de plus bas niveau) de la hiérarchie représentent les valeurs élémentaires. Les nœuds de la hiérarchie (non terminaux) sont les agrégats des données sur lesquels ont été appliqués la fonction « f ».

Supposons ici une dimension représentant les lieux de ventes. Ces lieux peuvent être agrégés en secteur, département puis région. La valeur de chaque

agrégat donnera la SOMME des valeurs prises par la VARIABLE pour ses descendants dans la hiérarchie.

On peut réaliser de cette manière l'équivalent d'un tableau croisé dynamique à N dimensions.

On pourrait démontrer que le nombre total de croisements entre les dimensions, et donc le nombre total de façons de voir une variable V, est égal à :

$$P * (P-1) * Np^2 / 2 \qquad (1)$$

où

P est le nombre de dimensions de l'entrepôt de données et

Np est le nombre moyen de modalités (c'est-à-dire de valeurs possibles) par dimension.

2) Architecture d'un système décisionnel

Les systèmes décisionnels classiques sont constitués de trois composants principaux :

a) l'entrepôt de données, qui fédère les données de l'entreprise, et qui offre un portail unique d'accès aux données au serveur de données (ici de type OLAP).

b) le serveur de données OLAP lui-même, qui organise et stocke les données sous la forme de tableaux croisés dynamiques de N dimensions.

c) enfin une série de postes clients.

Les données issues de l'entrepôt de données sont construites par agrégation en fonction d'indicateurs ou variables « métier ». En règle générale, la donnée élémentaire qui décrit un processus de vente pour un client, sa signalétique, etc., disparaît au profit de données décrivant des groupes homogènes, utiles pour l'analyse décisionnelle.

La base décisionnelle est conservée dans le serveur de données. Elle est accessible via des postes clients en environnement client/serveur classique ou via Internet/intranet.

Lors de l'ouverture d'une session cliente, le serveur de données OLAP ouvre un canal de transmission dédié entre le poste client et le serveur. Ce canal est constitué d'une zone tampon, conservée sur le serveur, zone appelée « espace de travail ». A chaque session, un nouvel espace de travail est créé. Cette gestion permet d'assurer l'accès simultané de plusieurs utilisateurs.

En outre, à chaque espace de travail est associé un identifiant unique de connexion.

L'espace de travail constitue le miroir des données visualisées sur le poste client. Toute interaction nécessitant le rafraîchissement des données provoque un appel au serveur via l'espace de travail.

Afin d'assurer l'intégrité des données contenues dans le serveur de données OLAP, une session n'a qu'un droit de lecture sur les données stockées dans le serveur, mais possède un droit de lecture et d'écriture dans l'espace de travail associé à cette session. Lors d'une déconnexion, l'espace de travail est détruit avec toutes les données qu'il contient.

3) Description globale des fonctionnalités de l'invention

L'invention a entre autres pour objectif de construire dynamiquement des filtres sur la ou les variables métier de l'utilisateur et de les insérer dans son interface - c'est-à-dire au niveau de son espace de travail - sans modifier le contenu applicatif du serveur. Ces filtres sont dans le présent exemple détruits lors de la déconnexion du client.

Plus précisément, soit une variable V telle que :

$$V = f(D1, D2, \ldots, DN)$$

Soit F* une nouvelle dimension, tel que

$$F^* = g(V^*)$$

avec

$V^* = f(D1, D2, \ldots, DP)$ et

(D1, D2, ..., DP) inclus dans (D1, D2, ..., DN)

alors il est possible de constituer une variable V' telle que

$$V' = f(D1, D2, \ldots, DN, F^*)$$

Avec V inclus dans V'

La dimension F* est construite par application d'une fonction « g » sur une autre agrégat de variables. La dimension F* est une dimension hiérarchique, comme on le verra en détail plus loin.

Ces fonctions sont réalisées directement dans l'espace de travail de la session, puis activées dans l'interface de l'utilisateur.

La variable V' est alors visualisable :

- Soit dans son contexte initial ; F* est alors agrégé sur le nœud de tête de sa hiérarchie (dans ce cas, V' est égale à la variable initiale V).
- Soit suivant les valeurs prises par F*, les valeurs F* étant utilisées pour filtrer la variable V.

4) Décomposition des fonctionnalités

Le procédé est réalisé par les étapes suivantes :

Etape 1 - Connexion, Extraction et Traitement

- récupération de l'identifiant de connexion de l'utilisateur et connexion à son espace de travail ;
- analyse de l'espace de travail pour identifier la variable métier que l'utilisateur cherche à exploiter ;
- extraction des données à analyser, à savoir des variables à discrétiser, ou des données à analyser ;
- traitements sur les données extraites.

Etape 2 - Mise à jour de l'espace de travail

- construction de la dimension F*, et de ses valeurs organisées sous forme simple ou hiérarchique ;
- création de la variable V' par la conjonction de V avec F*.

On notera ici que dans le cas où une dimension hiérarchique doit être construite, ceci est réalisé en associant à la dimension créée une relation de type « père – fils » entre les éléments qui la composent.

L'affectation des éléments de F* aux données décrivant V peut est réalisée par plusieurs méthodes :

- soit un pré-calcul des agrégats de données issues de l'interaction entre F* et les dimensions qui caractérisent la variable V (méthode dite de « ROLLUP ») ;
- soit la création d'un filtre qui calcul les agrégats en dynamique.

Etape 3 - Activation dans l'interface utilisateur

- affichage de la variable V' dans l'interface utilisateur.

Le procédé selon l'invention est récursif, c'est-à-dire il est possible de construire une variable V'' à partir de V', et ainsi de suite.

Le procédé selon l'invention a été testé et validé sur l'environnement décisionnel « OracleExpress » (Marque Déposée) commercialisé par la société américaine Oracle. Il peut bien entendu être mis en œuvre sur de nombreux autres environnements décisionnels du marché.

5) <u>Applications pratiques de la présente invention</u>

Le procédé et le système de l'invention permettent en particulier :

- de transformer une vision particulière d'une variable en une dimension et de la projeter sur la variable métier courante. La transformation d'une variable en dimension peut être réalisée par toute méthode de discrétisation telle que la méthode

dite de Fisher (ou « Fisherisation », qui consiste à découper de façon optimale d'une variable continue en N sous-ensembles) ;

- de construire une dimension qui correspond à une typologie de données, typologie issue d'une analyse de données sur la base OLAP, ou sur l'entrepôt de données associé au serveur OLAP, et d'en visualiser l'impact sur la variable courante.

Les applications sont alors multiples. Toutes applications décisionnelles de type OLAP ou réalisées sur des tableaux croisés dynamiques sont extensibles par ce procédé.

6) Exemples concrets

Soit une base de données OLAP de 1000 données, relative à la maintenance d'un certain matériel industriel et au coût de cette maintenance.

Les premières lignes de cette base de données sont indiquées sur la figure 1 des dessins.

Dans cette application, l'objectif du système décisionnel est de pouvoir étudier et optimiser la politique de maintenance et les coûts associés.

a) *Définition des variables métiers*

Les variables métiers sont :

V1 : Coût réel de la maintenance

V2 : Nombre cumulé de pannes en marche

Chaque variable est définie par les paramètres (nommés « dimensions » dans un système OLAP), avec ici six dimensions D1 à D6 qui sont définies comme suit :

D1 : nature du matériel

D2 : marque du matériel

D3 : utilisabilité, c'est-à-dire la vie restante du matériel (en %) de sa vie nominale ou préconisée

D4 : type de maintenance (ici, s'agit-il d'un remplacement, d'une rénovation ou d'un simple test de fonctionnement ?)

D5 : fréquence de maintenance requise

D6 : rapport (en %) entre coût de maintenance réel et coût de maintenance estimé

b) *Présentation des résultats*

Un système décisionnel connu peut fournir par exemple un tableau croisé dynamique de la façon suivante :

i) filtrage (présélection) de départ, effectuée indifféremment sur variables ou sur dimensions :

nature du matériel (D1):	tous
vie restante (D3) :	tous
type de maintenance (D4) :	tous
fréquence de maintenance (D5) :	tous
nombre de pannes en utilisation (V2) :	tous

ii) élaboration et présentation des données

Une telle présélection étant faite (dans le présent exemple, aucune), le système peut alors élaborer à partir d'agrégats pré-calculés de coûts de maintenance un tableau croisé dynamique exprimant par exemple le coût de maintenance en fonction des dimensions restantes, à savoir ici d'une part de la marque du matériel (dimension D2) et d'autre part du rapport entre coût réel et coût estimé (dimension D6), comme indiqué ci-dessous :

Somme des coûts réels	coût réel/est			
marque	80	100	120	Total
AAA	94615	320108	399719	814442
BBB	69277	378746	877090	1325113
CCC	113744	151948	221901	487593
Total	277636	850802	1498710	2627148

(On notera ici que les chiffres de coût de maintenance présentés ci-dessus ne correspondent pas aux données illustrées sur la figure 1, puisque celles-ci ne constituent qu'une partie de l'ensemble des données.)

Un tel tableau peut bien entendu, de façon conventionnelle, être accompagné de son graphique associé, tel qu'un histogramme (voir figure 2 des dessins).

La modification des valeurs d'une ou de plusieurs dimensions affichées permet de visualiser successivement les différents espaces de représentation de l'information. La variable « coût réel » étant toujours présentée dans les cellules centrales du tableau.

L'analyse du « coût réel » est réalisée par croisement 2 à 2 des valeurs des dimensions (ici D2 et D6).

Toujours dans le même exemple, si l'on se réfère à la formule (1) indiquée plus haut permettant de déterminer le nombre de possibilités de présentation d'une variable, on a ici 6 dimensions. En supposant que le nombre moyen de modalités dans les dimensions est égal à 5, il existe alors 375 manières possibles de présenter la variable « coût réel ».

c) *Apports pratiques de la présente invention*

Par rapport à cet existant, un objectif concret de la présente invention est de permettre de présenter les éléments caractéristiques d'une variable donnée en réduisant le nombre nécessaire de visualisations pour effectuer une analyse pertinente, et l'on va maintenant donner trois exemples de mise en œuvre de l'invention permettant d'atteindre un tel objectif :

- Exemple 1 : Transformation d'une variable en dimension
- Exemple 2 : Marquage des données en fonction d'une dimension
- Exemple 3 : Classification des données avec ou sans marquage des classes

c1) *Exemple 1 : Transformation d'une variable en dimension*

L'objectif de cet exemple est de montrer comment l'invention peut transformer une variable en dimension, puis analyser l'impact de cette dimension sur la variable métier courante. La transformation nécessite de discrétiser les valeurs de la variable sous la forme d'intervalles de valeurs.

Ce processus de discrétisation peut être fait de plusieurs manières connues, soit manuellement, soit par des méthodes de découpe optimale comme la méthode de Fisher simple (sans contrainte) ou généralisée (sous contraintes). Il est réalisé sur l'histogramme des valeurs de la variable considérée.

Chaque intervalle obtenu par ce processus de Fisher correspondra à une valeur d'une nouvelle dimension F* à visualiser dans l'OLAP. Une valeur supplémentaire sera rajoutée à la dimension F* qui correspondra à la somme des intervalles.

La relation des valeurs de F* avec les données de l'OLAP est effectuée soit par application d'un filtre sur la variable qui a été discrétisée, soit par un recalcul des valeurs des agrégats du tableau croisé (processus dit de « ROLLUP »). Le filtre ou le calcul des agrégats est donné par les bornes des intervalles extraites par la méthode de discrétisation.

En conservant l'exemple plus haut, on peut avoir par exemple le tableau croisé suivant :

i) filtrage (ici aucun)

nature du matériel (D1) :	tous
marque du matériel (D2) :	tous
vie restante (D3) :	tous
type de maintenance (D4) :	tous
fréquence de maintenance (D5) :	tous
nombre de pannes en utilisation (V2) :	tous

ii) élaboration et présentation d'un tableau exprimant, en fonction de la dimension D6 (coût réel/coût estimé), de première part le somme des coûts réels, de deuxième

part la somme des pannes en utilisation, et de troisième part le nombre total des matériels concernés (« effectif total »).

Coût réel/coût estimé	80	100	120	Total
Somme coût réel	277 636	850 802	1 498 710	2 627 148
Nombre de pannes en utilisation	11	99	173	283
Effectif total	116	353	434	903

On peut alors exécuter un processus de Fisher sur les agrégats de coûts réels figurant dans ce tableau, c'est-à-dire une séparation de l'intervalle des coûts réels en plusieurs sous-intervalles, par traitement sur l'histogramme des valeurs rencontrées. Dans le présent exemple, on élabore trois sous-intervalles correspondant respectivement à trois catégories de coûts individuels de maintenance : « faible », « moyen » et « fort ».

Un nouveau tableau croisé, ici à trois dimensions, peut alors être élaboré de la façon suivante :

i) présélection : comme ci-dessus

ii) élaboration et présentation du tableau

On présente ici, en fonction de deux dimensions principales, à savoir le rapport coût réel/coût estimé (D6) et une nouvelle dimension (F* dans les explications qui précèdent, et que l'on appellera D7 dans la suite), d'une part les agrégats de coûts réels (variable V1), de seconde part les agrégats de nombres de pannes en utilisation (variable V2), et de troisième part les effectifs concernés :

		Coût unitaire « Fisherisé » (D7)			
coût réel/est (D6)	Données	Faible	Moyen	Fort	Total
80	Somme coût réel (V1)	138 052	21 584	118 000	277 636
	Somme nombre de pannes en utilisation (V2)	10	1	0	11
	Somme effectif	98	4	14	116
100	Somme coût réel (V1)	201 680	649 122	0	
	Somme nombre de pannes en utilisation	50	49	0	99
	Somme effectif	128	225	0	353
120	Somme coût réel	319 836	808 938	369 936	1 498 710
	Somme nombre de pannes en utilisation	100	69	4	173
	Somme effectif	182	228	24	434
Total Somme coût réel		659 568	1 479 644	487 936	2 627 148
Total Somme nombre de pannes en utilisation		160	119	4	283
Total Somme effectif		408	457	38	903

Ainsi le processus de discrétisation par la méthode de Fisher, appliqué aux données d'entrepôt de données sans intervenir au niveau du serveur, mais seulement dans l'espace de travail concerné, permet à l'utilisateur, sans modifier nullement l'administration et la gestion de la base de données, d'avoir une vue différente - et surtout plus fine - de la manière dont se répartissent les coûts de maintenance.

c2) *Exemple 2 : Marquage des données en fonction de la variable fisherisée*

L'objectif est ici de caractériser les données en fonction du résultat du processus de Fisher appliqué à la variable « coût unitaire » des opérations de maintenance (comme décrit dans l'Exemple 1), afin de caractériser, selon un ou plusieurs groupes de critères, les populations d'individus concernés.

Cette caractérisation est réalisée ici par une méthode de calcul multivariée, appelée méthode des marquages, qui permet de trouver les paramètres discriminants.

Pour plus de détails quant à ce type de processus, connu en soi, on se référera par exemple à l'article de M. GETTLER-SUMMA, 1998 : Approche MGS:

Marquage et généralisation symbolique pour de nouvelles aides à l'interprétation en analyse de données - Cahier du CEREMADE (UMR 7534) N° 9830. Le point c21) plus loin traite plus en détail de ce processus de marquage.

Un tel processus, appliqué par exemple à la recherche de ce qui caractérise des matériels dont le coût de maintenance unitaire peut être qualifié de « fort », « moyen » ou « faible », va aboutir par exemple à des résultats du type :

- un coût unitaire fort est caractérisé à 92% par :

  * une population P1 constituée par les échangeurs de marque BBB (marquage No. 1) ;

  * une population P2 constituée par les échangeurs ayant une vie restante inférieure à 75 % (marquage No. 2).

A partir d'un tel résultat, on peut élaborer un nouveau tableau croisé :

i) filtrage

Le filtrage est ici double. On a en premier lieu :

nature du matériel (D1) :	échangeur
marque du matériel (D2):	BBB
vie restante (D3) :	tous
type de maintenance (D4) :	tous
fréquence de maintenance (D5) :	tous
nombre de pannes en utilisation (V2) :	tous
coût de maintenance unitaire (D7) :	fort

et en second lieu :

nature du matériel (D1) :	échangeur
marque du matériel (D2):	tous
vie restante (D3) :	< 75 %
type de maintenance (D4) :	tous
fréquence de maintenance (D5) :	tous

nombre de pannes en utilisation (V2) : tous

coût de maintenance unitaire (D7) : fort

ii) élaboration et présentation des données

On peut maintenant fabriquer le tableau suivant :

		coût réel/estimé		
Marquage	Données	80	120	Total
Echangeurs de marque BBB	Somme coût réel	36224	369936	406160
	Somme effectif	4	24	28
	Somme nombre de pannes en utilisation	0	4	4
Echangeurs de vie restante < 75	Somme coût réel	81776		81776
	Somme effectif	10		10
	Somme nombre de pannes en utilisation	0		0
Total Somme coût réel		118000	369936	487936
Total Somme effectif		14	24	38
Total Somme nombre de pannes en utilisation		0	4	4

Le tableau ci-dessus permet d'une part d'analyser la répartition du surcoût de maintenance, mais aussi de caractériser et d'identifier les individus provoquant ce surcoût.

(On observe ici qu'aucun des matériels filtrés ne possède de rapport coût réel/coût estimé égal à 100).

On va indiquer ci-dessous plus en détail d'une part ce en quoi consiste le processus de marquage, et comment il peut être appliqué à d'autres données que les données « Fisherisées », et d'autre part comment les marquages peuvent être pilotés selon un aspect de la présente invention.

c21) *Processus de marquage et généralisation*

Selon les enseignements de l'article de M. GETTLER-SUMMA indiqué plus haut, il est possible d'engendrer des marquages sur les données par l'analyse

détaillée de celles-ci de manière à trouver le plus petit sous-ensemble de descripteurs qui caractérisent la population étudiée, et de présenter les résultats sous la forme de requêtes OLAP.

Ainsi, supposons un jeu de données organisé sous la forme de N groupes. L'objectif est de pouvoir identifier dans chacun des groupes, les descripteurs qui caractérisent les sous populations de telle manière qu'il y ait :

- unicité dans les descriptions, ce qui revient à trouver les critères qui décrivent le groupe et uniquement celui là. ;
- recouvrement maximal de chaque groupe de données, ce qui revient à trouver les requêtes qui concernent le plus grands nombre d'individus ; et enfin
- erreur de recouvrement la plus faible, ce qui revient à rechercher des marquages qualitativement satisfaisants, c'est-à-dire recouvrant les individus et uniquement ceux-ci.

Dans l'exemple précédent, considérons la dimension D6 « coût réel/coût estimé ». Elle définit intrinsèquement trois groupes de population, à savoir :

- les matériels ayant un coût réel de maintenance de 120 %, soit 20% de plus que le coût estimé ou prévu ;
- les matériels de coût réel égal au coût estimé ;
- les matériels ayant un coût de maintenance plus faible (-20%) que le coût estimé.

L'objectif est de caractériser les individus de chaque classe et leurs descripteurs associés.

Ceci est illustré sur la figure 3 des dessins annexés, qui montre, dans le présent exemple, que la population d'individus dont la dimension D6 est égale à 80 % peut, par le processus précité, être subdivisé en trois classes ou sous-populations, à savoir :

- sous-population P1' constituée par les clapets de marque CCC ;
- sous-population P2' constituée par les échangeurs, dont le type de maintenance est « remplacement » et dont la fréquence de maintenance est le semestre ;
- sous-population P3' constituée par les appareils de mesure, dont le type de maintenance est « test » et dont la fréquence de maintenance est mensuelle.

(On observera ici que, contrairement à l'Exemple 2 plus haut, les populations sont disjointes, tout en couvrant l'essentiel de la population de départ).

On comprend ainsi que chaque classe est définie par les individus qui la composent (les petits rectangles sur la figure 3). Le marquage de chacun de ces individus dans l'espace de travail de l'utilisateur, permet de trouver les descriptions (critères ou filtres) des sous-populations de données ayant une très forte homogénéité.

Avantageusement, dans l'interface utilisateur, ces marquages s'expriment simplement par de courtes phrases qui correspondent étroitement aux valeurs prises par les descripteurs des données, et seulement à celles-ci.

Les marquages sont construits en tenant compte de l'organisation des données (structure hiérarchique, contraintes particulières sur les descripteurs comme les relations père-fils, identification des nœuds hiérarchiques les plus pertinents, etc.), l'objectif étant notamment de rechercher, si une bonne homogénéité est rencontrée à un niveau donné de la hiérarchie, de remonter d'un niveau dans celle-ci pour déterminer si l'homogénéité à ce nouveau niveau reste satisfaisante.

Les marquages obtenus sont en fait des requêtes de filtrage sur les données initiales, conservées en base de données et réutilisables en phase de pilotage (filtrage). Ils peuvent être engendrés soit pour identifier les individus soit selon une dimension existante, soit, dans l'exemple décrit ci-dessus, selon une dimension construite à cet effet notamment par un processus de Fisher ou par une classification automatique.

c22) *Pilotage par Marquage*

Le pilotage des données consiste à projeter le résultat des marquages sur une variable pour en mesurer l'impact. Dans le présent exemple, les marquages ont été construit sur les éléments signalétiques des pannes, puis les résultats ont été projetées sur la variable V1 correspondant au coût réel de maintenance.

Les résultats peuvent ainsi être synthétisés par des graphiques du genre de ceux illustrés sur les figures 4a et 4b des dessins. Ainsi la figure 4a illustre une

représentation prédéfinie au niveau du serveur, qui consiste à présenter les sommes des coûts individuels de maintenance selon la dimension D6 (rapport coût réel/coût estimé).

A partir de cette représentation (ou d'un tableau unidimensionnel associé), la présente invention permet d'élaborer un graphique tel qu'illustré sur la figure 4b, pour chaque barre de l'histogramme de la figure 4a (en l'espèce pour le cas où le paramètre selon la dimension D6 est 120%).

(On notera que dans cet exemple le processus de marquage a permis de trouver que c'étaient soit les matériels de type « appareil de mesure » dont le type de maintenance est « remplacement », soit les matériels de type « oscilloscope » dont la marque est « BBB » et dont le type de maintenance est « test », qui étaient représentatifs de cette catégorie particulière d'individus.

Selon une caractéristique avantageuse de la présente invention, on peut prévoir, via l'administrateur du serveur, d'étendre la base de données contenue dans le serveur pour y inclure des paramètres de marquage.

Ces marquages stockées dans la base de données peuvent ainsi être ré-utilisés pour en suivre leur l'évolution, comme le montrent en détail les figures 5a et 5b des dessins.

Ainsi la figure 5a est une représentation prédéfinie de la répartition des coûts de maintenance, par un graphique bidimensionnel de type histogramme empilé, d'une part selon une dimension additionnelle D8 (temps, par exemple exprimé en mois) qui dans cet exemple existe dans la base de données (abscisse), et d'autre part selon la dimension D6, ce graphique pouvant être fourni en standard par le serveur.

Grâce au processus de marquage précité, on peut maintenant réaliser un graphique bidimensionnel tel que celui illustré sur la figure 5b (ici de type aires empilées), avec une abscisse correspondant ici encore à la dimension D8, et une ordonnée correspondant à une nouvelle dimension D9 dont les paramètres sont les marquages indiqués ci-dessus.

Il devient ainsi possible de suivre l'évolution dans le temps des sous-populations ou classes qui, à l'origine, ont été marquées comme représentatives de la valeur du paramètre en question (ici coût réel/coût estimé = 120%).

Il est intéressant d'observer ici que les individus résiduels occupant la sous-population « RESTE » (zone supérieure dans les graphiques des figures 4b et 5b), appelée aussi résidu, peuvent faire à tout moment l'objet d'un marquage spécifique, marquage qui viendra compléter les marquages existants. Par ailleurs, une population identifiée à un instant t, peut aussi faire l'objet d'un marquage complet.

c3) *Exemple 3 : Classification des données*

Selon une autre possibilité de la présente invention, on peut également chercher à identifier le meilleur classement pour les données en s'appuyant sur des techniques de classification automatique.

Un premier objectif est de trouver la meilleure typologie de classification. Le résultat obtenu est alors un partitionnement optimal dans lequel tous les individus sont répartis. Chaque individu de la population se trouve alors affecté à une partition et une seule. On notera ici que de nombreux outils d'analyse de données existant sur le marché fournissent ce type d'algorithmes.

En reprenant l'exemple précédent, on peut obtenir une partition optimale en 3 classes désignées par Classe 1, Classe 2 et Classe 3. Le tableau croisé peut alors être modifié pour tenir compte de cette répartition :

i) filtrage

nature du matériel (D1) :	tous
marque du matériel (D2) :	tous
vie restante (D3)	tous
type de maintenance (D4) :	tous
fréquence de maintenance (D5) :	tous
Fisher :	tous

ii) élaboration et présentation des données

On réalise ici un tableau dont une dimension est le rapport coût réel/coût estimé (D6) et dont l'autre dimension est la classification obtenue.

		Classification automatique			
réel/estimé	Données	Classe 1	Classe 2	Classe 3	Total
80	Somme effectif	116			116
	Somme coût_réel	277 636			277 636
	Somme Nb_panne	11			11
100	Somme effectif			353	353
	Somme coût_réel			850 802	850 802
	Somme Nb_panne			99	99
120	Somme effectif	129	50	360	539
	Somme coût_réel	2 026 836	133 626	995 148	3 155 610
	Somme Nombre de pannes	4	48	121	173

Une nouvelle dimension F* (notée ici D10) est ainsi construite avec le résultat de la classification automatique.

Ce résultat peut être étendu en identifiant par la technique des marquages des types d'individus associés à chacune des classes. Les marquages, organisés par classe, sont rajoutés sur la dimension D10, qui devient alors une dimension hiérarchique :

On obtient la nouvelle dimension suivante :

Classe 1 : couverture des marquages de 77%

1er Marquage classe 1 : échangeur

2ème Marquage classe 1 : clapet, marque CCC

3ème Marquage classe 1: matériel, neuf, marque CCC, testé

(on entend ici par « neuf » un paramètre vie restante de 100%)

Classe 2 : couverture des marquages de 0%

néant

Classe 3 : couverture des marquages de 82 %

1er Marquage classe 3 :	matériel, marque AAA, testé
2ème Marquage classe 3 :	matériel, marque BBB, maintenu par mois ou par trimestre ou rénové

L'exemple ci-dessus montre que, lorsque les marquages sont générés sur un sous-ensemble des descripteurs pris en compte dans la classification, il se peut qu'il n'y ait aucun marquage caractéristique pour une classe donnée (ici pour la Classe 2).

Ici encore, chaque marquage décrit exclusivement les individus de sa classe.

Comme pour les autres dimensions, les nœuds « Classe » sont les agrégats de la dimension F* sur lesquels est appliquée une fonction f qui définit la variable métier courante de l'utilisateur.

Dans ce cas, la dimension F* est hiérarchique, et on lui associe à cet effet une relation qui relie les éléments de la hiérarchie entre eux.

Chaque marquage de classe est relié aux données du système décisionnel en appliquant aux données la requête de filtrage correspondant au résultat du processus de marquage.

On obtient ainsi, comme indiqué plus haut, la nouvelle variable V' par application de f sur les dimensions décrivant V et F* :

$$V' = f(D1, D2, \ldots, DN, F^*)$$

et ici encore, le calcul de V' par la fonction f est réalisé par recalcul des agrégats intermédiaires, à l'aide d'un processus de type « ROLLUP ».

Bien entendu, la présente invention n'est nullement limitée aux exemples ci-dessus, mais l'homme du métier saura y apporter de nombreuses variantes ou modifications.

REVENDICATIONS

1. Système de traitement de données, comprenant :

- un serveur contenant dans une mémoire un jeu de données individuelles organisées sous forme d'au moins une variable (V1) à N dimensions (D1-D6), chaque donnée individuelle contenant au moins une variable individuelle en association avec des valeurs prédéfinies de paramètres selon les N dimensions, et

- au moins un poste client communiquant avec le serveur via un espace de travail associé au serveur,

le serveur étant apte à placer dans ledit espace de travail des données constituées par des combinaisons prédéfinies (V = f(D1, ..., D6)) desdites variables individuelles à partir d'un filtrage sur lesdits paramètres, réalisé selon au moins une dimension choisie par un utilisateur parmi les N dimensions,

système caractérisé en ce qu'il comprend également, à l'extérieur du serveur, des moyens pour constituer temporairement des paramètres additionnels (F*) à partir d'un traitement d'analyse sur des combinaisons choisies de variables placées dans l'espace de travail, et des moyens pour élaborer dans ledit espace de travail de nouvelles combinaisons desdites variables (V' = f (D1, D2, ..., DN, F*)) à partir d'un nouveau filtrage mettant en jeu lesdits paramètres additionnels.

2. Système selon la revendication 1, caractérisé en ce que les moyens pour constituer des paramètres additionnels comprennent des moyens pour établir des catégories selon un processus de catégorisation sur lesdites combinaisons de variables, lesdits paramètres additionnels étant constitués par lesdites catégories.

3. Système selon l'une des revendications 1 et 2, caractérisé en ce que les moyens pour constituer des paramètres additionnels comprennent des moyens pour établir des catégories selon un processus de marquage sur lesdites combinaisons de variables, lesdits paramètres additionnels étant constitués par des combinaisons de paramètres existants à partir desquelles un filtrage sur les données est effectué.

4. Système selon l'une des revendications 1 à 4, caractérisé en ce que les moyens pour élaborer de nouvelles combinaisons (V' = f (D1, D2, ..., DN, F*)) desdites variables comprennent des moyens pour pré-calculer et stocker lesdites nouvelles combinaisons.

5. Système selon l'une des revendications 1 à 3, caractérisé en ce que les moyens pour élaborer de nouvelles combinaisons desdites variables comprennent des moyens pour calculer dynamiquement lesdites nouvelles combinaisons (V' = f (D1, D2, ..., DN, F*)) à partir d'un filtrage appliqué au serveur et mettant en jeu lesdits paramètres additionnels.

6. Utilisation d'un système selon l'une des revendications 1 à 5 dans un système décisionnel par traitement analytique en ligne (OLAP).

7. Procédé de traitement de données et de visualisation desdites données sur un poste client à partir d'un serveur contenant dans une mémoire un jeu de données individuelles organisées sous forme d'au moins une variable (V1, V2) à N dimensions, chaque donnée individuelle contenant au moins une variable individuelle en association avec des valeurs prédéfinies de paramètres selon les N dimensions (D1-D6), caractérisé en ce qu'il comprend les étapes consistant à :

- fournir au poste client, via un espace de travail associé au serveur, des données constituées par des combinaisons prédéfinies (V = f (D1, ..., DN)) desdites variables individuelles à partir d'un filtrage sur lesdits paramètres, réalisé selon au moins une dimension choisie par un utilisateur parmi les N dimensions,

- dans l'espace de travail, constituer temporairement des paramètres additionnels à partir d'un traitement d'analyse effectué sur des combinaisons choisies desdites variables, et

- dans l'espace de travail, élaborer de nouvelles combinaisons (V' = f (D1, D2, ..., DN, F*)) desdites variables à partir d'un nouveau filtrage mettant en jeu lesdits paramètres additionnels.

8. Procédé selon la revendication 7, caractérisé en ce que ledit traitement d'analyse est un traitement de catégorisation sur lesdites combinaisons de variables, lesdits paramètres additionnels étant constitués par lesdites catégories.

9. Procédé selon l'une des revendications 7 et 8, caractérisé en ce que ledit traitement d'analyse comprend un processus de marquage, lesdits paramètres additionnels étant constitués par des combinaisons de paramètres existants à partir desquelles un filtrage sur les données du serveur est effectué.

10. Procédé selon l'une des revendications 7 à 9, caractérisé en ce que l'étape d'élaboration de nouvelles combinaisons desdites variables comprend un pré-calcul et un stockage desdites combinaisons dans l'espace de travail.

11. Procédé selon l'une des revendications 7 à 10, caractérisé en ce que l'étape d'élaboration de nouvelles combinaisons desdites variables comprend un calcul dynamique desdites combinaisons à partir d'un filtrage sur les données du serveur mettant en jeu lesdits paramètres additionnels.

12. Procédé selon l'une des revendications 7 à 11, caractérisé en ce que lesdites données appartiennent à un système décisionnel par traitement analytique en ligne.

D1	D2	D3	D4	D5	D6	V1	V2
Matériel	*marque*	*utisabilité(%)*	*type maintenance*	*fréquence cout : réel / estimé*		*coût réel*	*Nb panne en marche*
Réflectomètres	BBB	25	remplacement	trimestre	100	11320	2
échangeur	AAA	25	remplacement	semestre	80	22776	0
clapet	CCC	100	rénovation	mois	80	1532	0
filtres	AAA	25	remplacement	trimestre	120	875	1
Ponts de mesure	BBB	25	remplacement	mois	100	7118	1
Oscilloscopes	AAA	25	remplacement	mois	100	3830	1
clapet	BBB	50	remplacement	semestre	120	2298	0
échangeur	BBB	25	test	trimestre	120	6996	2
échangeur	BBB	100	test	trimestre	120	78900	1
Oscilloscopes	BBB	25	remplacement	semestre	100	7118	1
Ponts de mesure	BBB	25	remplacement	mois	100	7118	1
Ponts de mesure	AAA	50	remplacement	trimestre	100	16438	1
Ponts de mesure	AAA	75	remplacement	trimestre	100	7118	1
Ponts de mesure	AAA	25	remplacement	mois	100	3830	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

FIG_1

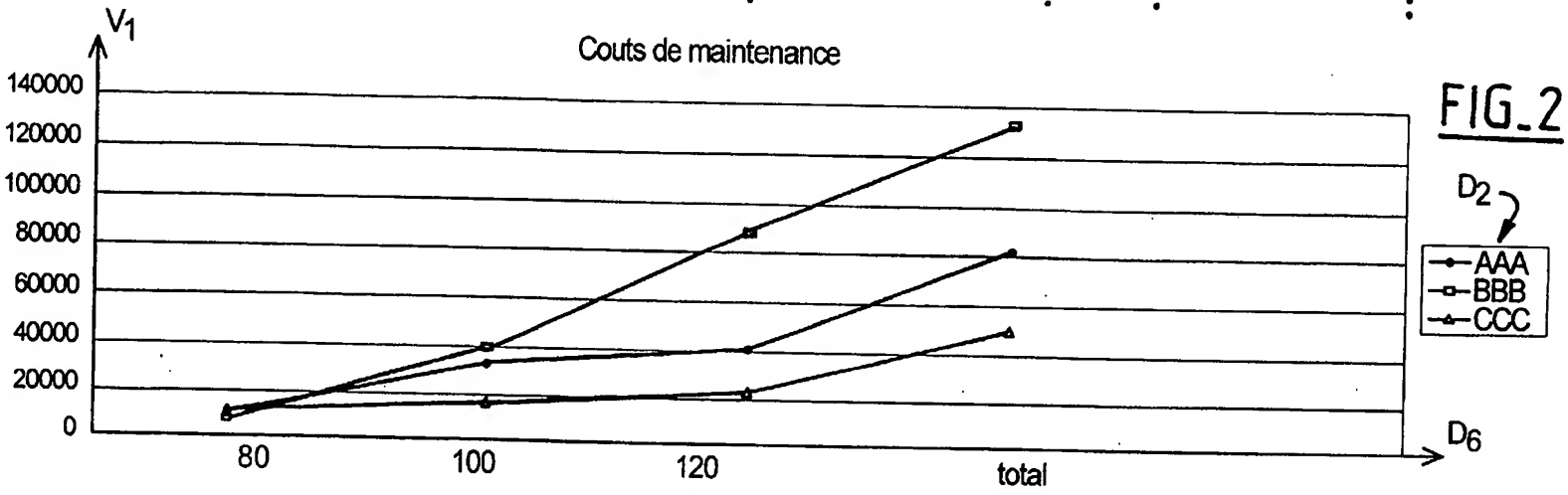


FIG_2

2799024

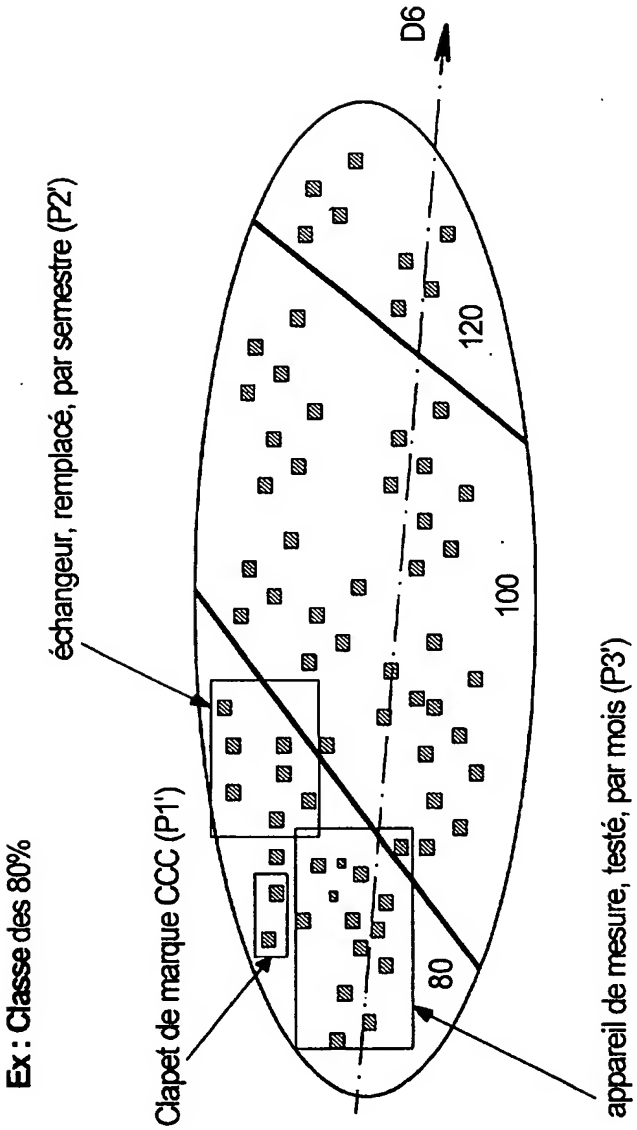
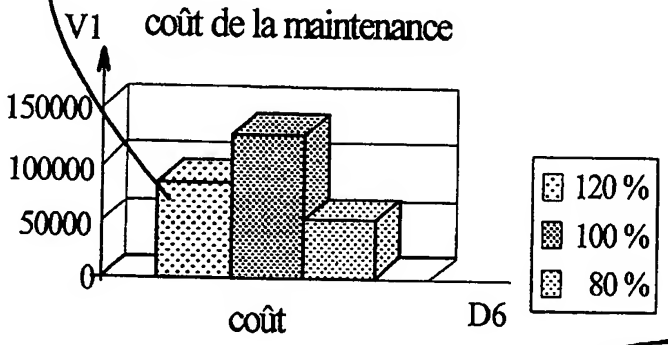

D6
120
100
80
échangeur, remplacé, par semestre (P2')
Clapet de marque CCC (P1')
appareil de mesure, testé, par mois (P3')
Ex : Classe des 80%

FIG.3

FIG_4a



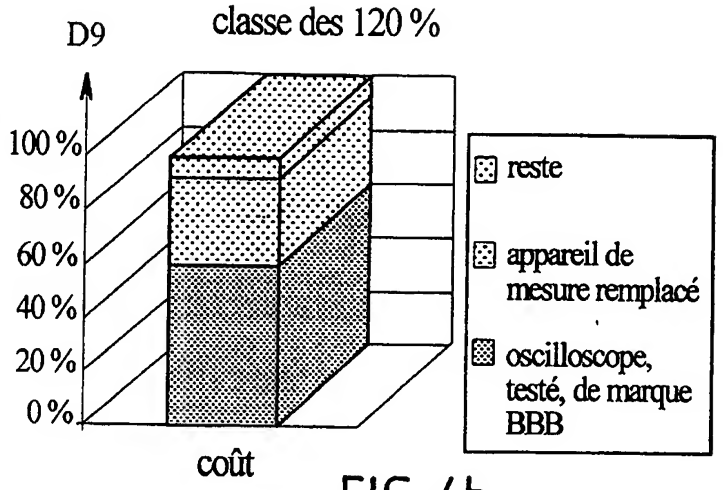
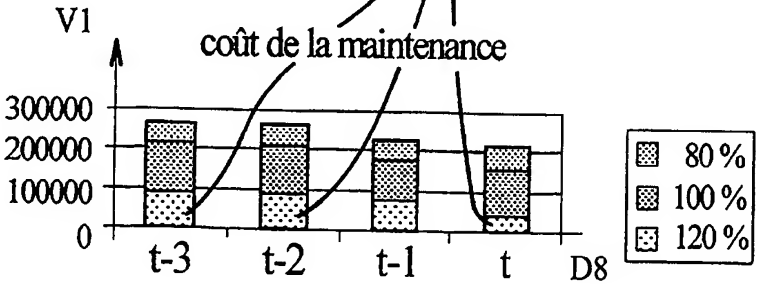


FIG_4b

FIG_5a



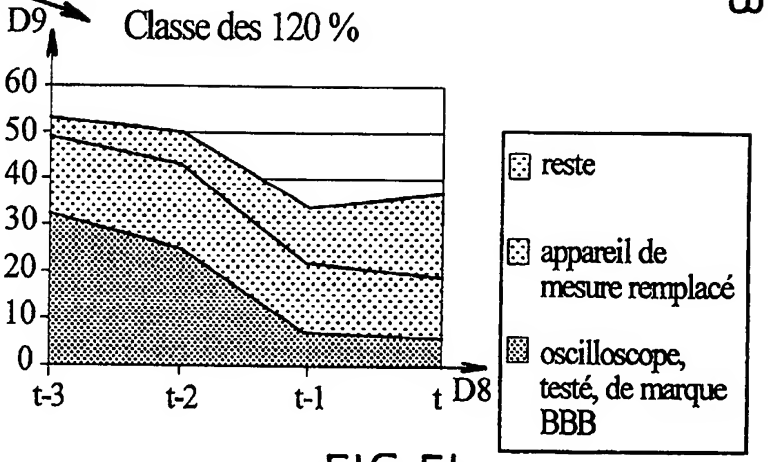


FIG_5b

RÉPUBLIQUE FRANÇAISE

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

2799024
N° d'enregistrement national

FA 584413
FR 9912141

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	CHAUDHURI S ET AL: "An overview of data warehousing and OLAP technology" SIGMOD RECORD,US,ASSOCIATION FOR COMPUTING MACHINERY, NEW YORK, vol. 26, no. 1, mars 1997 (1997-03), pages 65-74-74, XP002115173 * page 68, colonne de droite, ligne 6 - page 69, colonne de droite, ligne 7; figure 2 *	1-12	G06F17/30
A	US 5 926 818 A (MALLOY WILLIAM EARL) 20 juillet 1999 (1999-07-20) * colonne 10, ligne 18 - ligne 55; figure 1 *	1,7	
A	FLOHR U: "OLAP BY WEB" BYTE,US,MCGRAW-HILL INC. ST PETERBOROUGH, vol. 22, no. 9, 1 septembre 1997 (1997-09-01), pages 81-84, XP000726368 ISSN: 0360-5280 * page 81, colonne de gauche, ligne 1 - page 83, colonne de droite, ligne 54 *	1,7	
			DOMAINES TECHNIQUES RECHERCHÉS (Int.Cl.7)
			G06F

Date d'achèvement de la recherche	Examinateur
30 juin 2000	Deane, E

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 12.99 (P04C14)